Universiteit Leiden
Governance and Global Affairs

udp UNIVERSIDAD DIEGO PORTALES

tDl

Benchmarking LLMs in Political Content Text-Annotation

Proof-of-Concept with Toxicity and Incivility Data

Bastián González-Bustamante
Leiden University & Universidad Diego Portales
✉ b.a.gonzalez.bustamante@fgga.leidenuniv.nl

Presentation delivered at the 8th Monash-Warwick-Zurich Text-as-Data Workshop
September 16-17, 2024

Table of Contents

Introduction

Introduction

This article benchmarked the ability of a number of models, such as **Google's Perspective algorithm**, property or closed-source **OpenAI's GPTs** and **open-source LLMs**, to perform annotation tasks on political content.

The models were deployed through API or locally to identify **toxicity and incivility on social media** on a novel protest event dataset comprising more than 3M digital interactions.

- It offers insights into the performance of different GPTs and open-source LLMs in annotation tasks.

Related Work

Text-as-Data and Perspective API

The text-as-data approach appears as a relevant cornerstone due to the limited number of texts that human coders can annotate (Schmidt et al., 2024).

The **Perspective API**, developed by Jigsaw and Google, is one of the top-shelf options for classifying toxicity and incivility. It was trained on a variety of other sources labelled by crowdsource raters using distilled BERT models.

In academic research, the algorithm has been used to detect toxicity and uncivil comments on Twitter (Hopp et al., 2020; Orchard et al., 2024; Schmidt et al., 2024), Facebook (Hopp et al., 2020; Kim et al., 2021; Schmidt et al., 2024), news comments (Orchard et al., 2024; Schmidt et al., 2024) and Wikipedia content (Pavlopoulos et al., 2020).

Open-source LLMs

Researchers have been using GPTs through OpenAI's API for a variety of tasks (Gilardi et al., 2023; He et al., 2024), and despite different **concerns related to reproducibility, privacy and openness** of this pay-per-use form, this way tends to offer resources beyond those usually available to the average researcher in social sciences (easy to deploy).

Recent studies have shown that some open-source LLMs are able to perform or even surpass GPT-3.5 and 4 in zero-shot tasks, only lagging in fine-tuned GPTs (Alizadeh et al., 2024). In addition, some "small" open-source models, $\leq 7B$ parameters, thanks to quantisation techniques, are able to perform well and offer a considerable cost reduction (Irugalbandara et al., 2024, see also Bucher and Martini, 2024).

Data and Methods

Data and Gold Standard

Novel dataset on protest events

Protests against the coronavirus and judicial reform (August 2020) ✔

Social outburst in Chile (October 2019) ✔

Protests against education budget cuts (May 2019) ✖

∼3.5M messages in total

* ∼5M including Brazil

LLMs zero-shot classification

› Random, balanced sample of **1,000 messages**

Open-source LLMs deployed locally

› Minimum temperature to ensure reproducibility**

Intel Core i9-14900K CPU

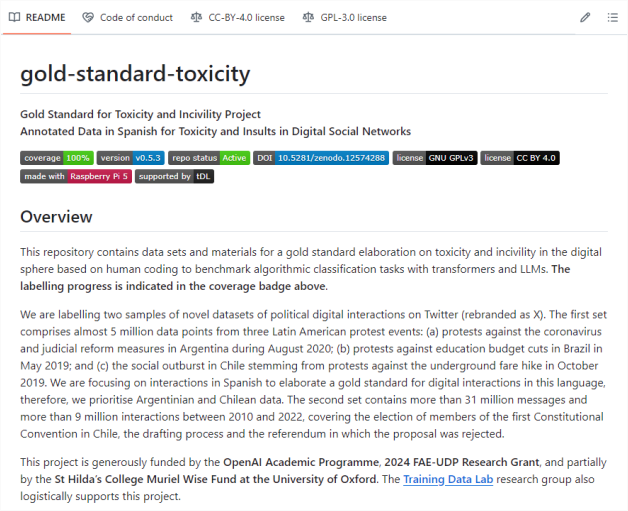
NVIDIA GeForce RTX 4070 Super Windforce OC 12GB GPU

64 GB of RAM (DDR5-4800)

Windows Subsystem for Linux v2.1.5.0 and Ollama v0.1.44 and v0.3.10

** We ran some temperature experiments to test models' hallucinations and reproducibility.

Gold Standard



README | Code of conduct | CC-BY-4.0 license | GPL-3.0 license

gold-standard-toxicity

Gold Standard for Toxicity and Incivility Project
Annotated Data in Spanish for Toxicity and Insults in Digital Social Networks

coverage 100% | version v0.5.3 | repo status Active | DOI 10.5281/zenodo.12574288 | license GNU GPLv3 | license CC BY 4.0
made with Raspberry Pi 5 | supported by tDL

Overview

This repository contains data sets and materials for a gold standard elaboration on toxicity and incivility in the digital sphere based on human coding to benchmark algorithmic classification tasks with transformers and LLMs. **The labelling progress is indicated in the coverage badge above.**

We are labelling two samples of novel datasets of political digital interactions on Twitter (rebranded as X). The first set comprises almost 5 million data points from three Latin American protest events: (a) protests against the coronavirus and judicial reform measures in Argentina during August 2020; (b) protests against education budget cuts in Brazil in May 2019; and (c) the social outburst in Chile stemming from protests against the underground fare hike in October 2019. We are focusing on interactions in Spanish to elaborate a gold standard for digital interactions in this language, therefore, we prioritise Argentinian and Chilean data. The second set contains more than 31 million messages and more than 9 million interactions between 2010 and 2022, covering the election of members of the first Constitutional Convention in Chile, the drafting process and the referendum in which the proposal was rejected.

This project is generously funded by the **OpenAI Academic Programme**, **2024 FAE-UDP Research Grant**, and partially by the **St Hilda's College Muriel Wise Fund at the University of Oxford**. The Training Data Lab research group also logistically supports this project.

The sample was manually annotated by a team of five high-skilled human coders on Labelbox.

Each message was hand-coded by two different coders of the team, therefore, we have **two ground-truth labels for each sample observation**.

GitHub repository.

Gold Standard

We used a binary codification performed using the core definitions of Perspective for toxicity as follows:

Human annotation

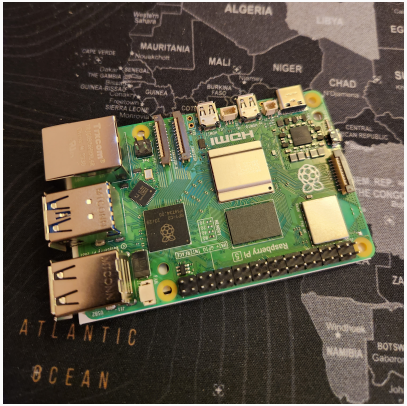
Classify the category of the comment as either **TOXIC** or **NONTOXIC**. **TOXIC:** Rude, disrespectful, or unreasonable comments that are likely to make someone leave the discussion or stop sharing their perspective. **NONTOXIC:** Civil or nice comments that are unlikely to discourage conversation.

- Krippendorff's α (bootstrap 1,000x) 0.944 (95% CI [0.919, 0.961]).
- The process involved 7.2 hours of annotation and 1.2 hours of revision.

Perspective API Classification on Raspberry Pi 5

First, we ran Perspective API. The classification implied **2,411 hours** of computing **(entire dataset, not only the sample)**.

The process was deployed on a **Raspberry Pi 5**, a small computer board with an ARM CPU and meagre power consumption. Our footprint using Raspberry Pi 5 was 6.51 kg of CO2, which implied a **reduction of 96% in emissions** that we would have generated running the tasks on an average cloud computing service.



* Picture of the Raspberry Pi 5 used for classification.

GPTs and Open-Source LLMs for Text-Annotation

OpenAI's GPTs

GPT-4o, GPT-4o mini, GPT-4, GPT-4 Turbo and GPT-3.5 Turbo

* o1-preview and o1-mini, released recently on September 12, 2024, are not fully available to all API users.

SOTA open-source LLMs

Llama 3.1, Hermes 3, Gemma 2 9B, Gemma 2 27B and Mistral NeMo

Slightly outdated open-source LLMs

Nous Hermes 2 Mixtral, Nous Hermes 2, Mistral OpenOrca, Orca 2 and Aya

Prompt Strategy

Our prompt strategy was the very **same used on the human annotation** based on the core definitions of Perspective and Google for toxicity with the following system message:

System message

Classify the category of the comment as either **TOXIC** or **NONTOXIC**. **TOXIC:** Rude, disrespectful, or unreasonable comments that are likely to make someone leave the discussion or stop sharing their perspective. **NONTOXIC:** Civil or nice comments that are unlikely to discourage conversation.

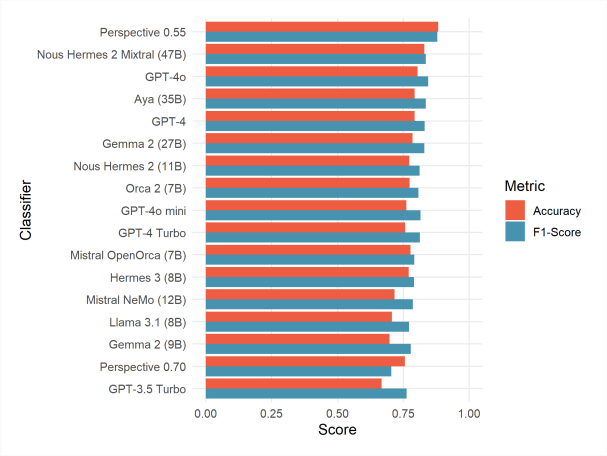
Along with providing texts of our balanced sample, we also listed the categories for the task as follows: **"Respond with only the category (TOXIC or NONTOXIC). Do not provide any additional analysis or explanation."**

Results

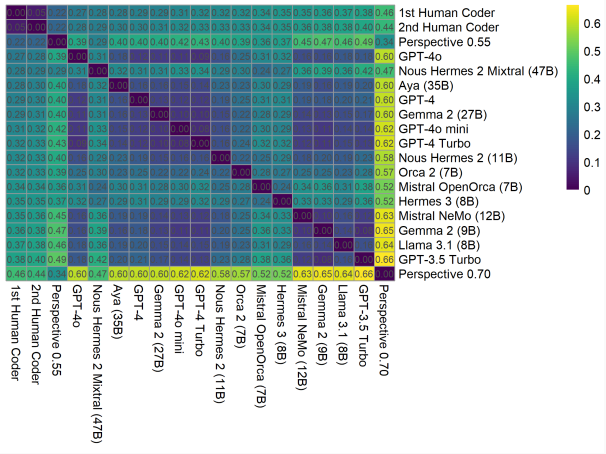
Error Rate Analysis

Perspective at 0.55 (relaxed threshold), **Nous Hermes 2 Mixtral** 47B, and **GPT-4o** show the highest F1-score in classifying toxicity considering our gold standard.

Llama 3.1, the novel Meta's SOTA, shows one of the lowest performances (still acceptable). However, this is the lightest version since 70 and 405B are beyond our computing infrastructure.



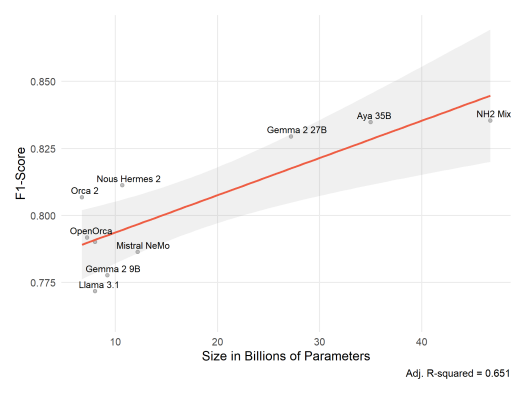
Jaccard Distance between Gold Standard and Classifiers

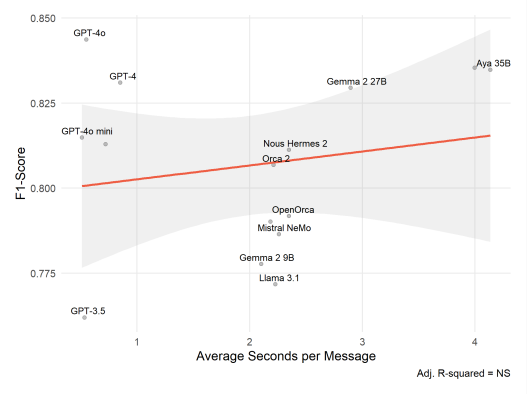


With an intermediate performance, **Mistral OpenOrca** and **Hermes 3**, both with parameters between 7 and 8B, show indexes around 0.30, therefore, they are not highly similar to other models. This could open avenues for ensemble annotations or stacking.

The same applies to Nous Hermes 2 Mixtral, which showed one of the highest performances.

Performance, Parameters and Computing Time





Computing Time vs Reproducibility

Time per obs.

Human coder: 15.125s

Perspective API: 1.175s

OpenAI's GPTs: 0.633s

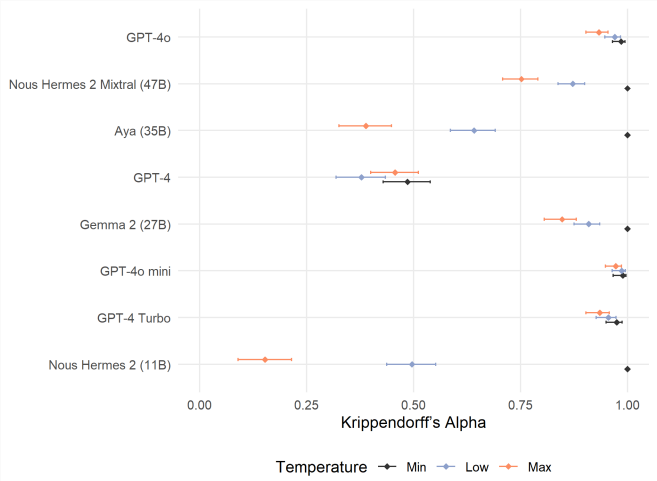
Open-source LLMs: 2.672s

How much faster than humans?

Open-source LLMs ∼6x

Perspective API ∼13x

OpenAI's GPTs ∼24x



Takeaways

Takeaways

☑ **Perspective API** using a laxer threshold, **GPT-4o**, and **Nous Hermes 2 Mixtral** outperform other LLM's zero-shot classification annotation.

☑ Nous Hermes 2 and Mistral OpenOrca, with a smaller number of parameters, are able to perform the task with high performance, being attractive options that could offer **good trade-offs** between performance, implementing costs and computing time.

☑ Although GPTs tend to show not only excellent computing time but also overall good levels of reliability, **only open-source LLMs ensure full reproducibility** in the annotation.

☑ **New avenues.** More complex classifications (few-shot or chain-of-thought); and use this to analyse incivility dynamics on specific cases.

Acknowledgements and Funding

This work was supported by the OpenAI's Academic Programme and the Faculty of Administration and Economics at the Universidad Diego Portales, Chile. I also thank the Institute of Security and Global Affairs at Leiden University, Netherlands, and the Training Data Lab research group for their support.

udp UNIVERSIDAD DIEGO PORTALES

Universiteit Leiden
Institute of Security and Global Affairs

References

References I

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., and Gilardi, F. (2024). Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning. Preprint, arXiv.

Bucher, M. J. J. and Martini, M. (2024). Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. Preprint, arXiv.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., and Chen, W. (2024). AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. Preprint, arXiv.

Hopp, T., Vargo, C. J., Dixon, L., and Thain, N. (2020). Correlating Self-Report and Trace Data Measures of Incivility: A Proof of Concept. *Social Science Computer Review*, 38(5):584–599.

References II

Irugalbandara, C., Mahendra, A., Daynauth, R., Arachchige, T. K., Dantanarayana, J., Flautner, K., Tang, L., Kang, Y., and Mars, J. (2024). Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs in Production. Preprint, arXiv.

Kim, J. W., Guess, A., Nyhan, B., and Reifler, J. (2021). The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication*, 71(6):922–946.

Orchard, X., González-Bustamante, B., and Cisternas, C. (2024). Amplification, contestation, and hostile engagement with news in social media: Comparing interactions with journalistic content before and after a protest cycle. In *Media and Governance in Latin America, Spain and Portugal*, Valencia, Spain. Universitat de València.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., and Androutsopoulos, I. (2020). Toxicity Detection: Does Context Really Matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Schmidt, F., Stier, S., and Otto, L. (2024). Incivility in Comparison: How Context, Content, and Personal Characteristics Predict Exposure to Uncivil Content. *Social Science Computer Review*. OnlineFirst.

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com

Presentation compiled with LaTeX and some

Download the latest version from GitHub